Specification

A METHOD OF FABRICATING A CMOS DEVICE WITH DUAL METAL GATE ELECTRODES

BACKGROUND OF THE INVENTION

The present application claims priority from United States provisional patent application Serial No. 60/464,936 filed April 22, 2003.

FIELD OF THE INVENTION

The present invention relates generally to methods of fabricating CMOS devices and more particularly to a method of fabricating a CMOS device with dual metal gate electrodes using a consumable thin buffer layer between the metal gate and the gate dielectric.

BRIEF DESCRIPTION OF THE PRIOR ART

In the construction of CMOS devices, doped polysilicon is commonly used for gate electrodes. Polysilicon is convenient because it can be doped to achieved the desired work functions in the two CMOS GATES. However, problems arise as the CMOS device is scaled to smaller dimensions. High resistivity, reduced inversion charge density and transconductance, and undesirable depletion of doped polysilicon gate electrodes can occur, resulting in a detrimental increase in the thickness of the gate oxide layers. There is also a problem with boron penetration by diffusion from the polysilicon into thin gate oxide layers.

As the equivalent gate oxide thickness (EOT) decreases below 1.0nm, the capacitance connected with the depletion layer in the polysilicon gate becomes an important limiting factor in EOT scaling. Therefore, it is projected that a metal gate and in particular a dual metal gate may be required in 50nm and smaller gate lengths. In the dual metal gate, an N-MOS metal and a P-MOS metal are used for the dual gates. However, current dual metal

gate technology has unsolved problems in process integration, especially in the procedure of lithographically masking and removing the first metal of the dual metal gates deposited on a portion of a wafer without generating etching damage to the gate dielectric. The usual method for fabricating dual metal gate electrodes is to deposit the first metal on top of the gate dielectric. The first metal is then removed by lithographically masking and selective etching from one of the well regions, which may be the n-well or p-well region. After that the second metal is deposited on top of the first metal as well as the exposed dielectric. Unfortunately the etching chemical solutions can also attack and remove a portion of the gate dielectric. This is a practical obstacle in the use of dual metal gate technology in production, even if the right metals are successfully identified. Another method to fabricate dual metal gate electrodes involves the use of ion implantation technology. In this case, a metal is deposited on top of the gate dielectrics, and one of the well regions is covered with photoresist. Ion implantation is then applied to one of the metal electrodes, which changes the work function of the metal. As a result, two different work functions of metal electrodes are obtained. However, the ion implantation can damage the gate dielectric, resulting in degradation of gate dielectric performance.

SUMMARY

Briefly, a preferred embodiment of the present invention includes a method of constructing a dual metal gate CMOS structure that uses an ultra thin aluminum nitride (AlN_x) buffer layer with a thickness typically less than 20nm. The layer lies between the metal gate and gate dielectric during processing for protecting the gate dielectric during the metal gate etching process. After dual metal gates are formed, the CMOS structure is subjected to an annealing temperature. During the annealing, the buffer layer is completely consumed through reaction with the metal gate and new metal alloys are formed that have optimal work functions . The annealing process causes only a minimal increase in the equivalent thickness.

IN THE DRAWING

Fig. 1 is a flow chart for describing the method of the present invention;

Fig. 2A illustrates the deposition of a buffer layer over a gate dielectric;

Fig. 2B illustrates deposition of a first metal over the buffer layer, and preparation for removing the first metal from a portion of the buffer layer;

Fig. 2C shows the structure with the unwanted portion of first metal removed;

Fig. 2D illustrates deposition of a second metal;

Fid. 2E illustrates preparation for selective etching of the first and second metals and the buffer layer;

Fig. 2F shows the CMOS structure with the metal removed as mentioned in reference to Fig. 2E;

Fig. 2G illustrates annealing and the resultant effect on the buffer layer;

Fig. 3 is a table showing the etching rates of various films;

Fig. 4 is a graph of capacitance versus gate voltage for various platinum (Pt) gates after anneal;

Fig. 5 is a graph of capacitance versus gate voltage for Hf-AlN_x/S_iO_2 and Ta-AlN_x/S_iO_2 gate after anneal;

Fig. 6 is a graph of forward gate voltage versus the thickness of the oxide region after anneal, for Ta-AlN_x and Hf-AlN_x gate metals; and

Fig. 7 is a graph of resultant equivalent gate oxide thickness variation as a function of annealing temperature for two different thicknesses of AlN_x buffer layers.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

The method of the present invention is illustrated in the flow chart of Fig. 1. The process/method fabricates dual metal gate electrodes of a CMOS device. This method allows each metal gate electrode to have its optimal work function, namely 4.4v for NMOS and 4.9v for PMOS. The process begins (block 10) with forming a gate dielectric 26 on a substrate 22 which has a p-well region and n-well region for use in the formation of NMOSFET and PMOSFET devices in a CMOS. In a preferred embodiment the gate dielectric is thermally grown S_iO_2. A buffer layer 28 is then deposited (block 11) on the gate dielectric 26 of the CMOS p and n-well structures. Methods for depositing the buffer layer include physical vapor deposition (PVD), chemical vapor deposition (CVD), atomic layer deposition (ALD), and sputtering. The buffer layer material is selected to meet three criteria: (a) It must be chemically resistant to protect the underlying gate dielectric from exposure to an etchant used during etching of a gate metal that is deposited on the buffer; (b) It must be consumable during annealing so as to form dual metal alloys through reaction with the gate metals during an annealing procedure so that the consumed buffer layer does not increase the final equivalent oxide thickness; and (c) The buffer layer material must be selected so that the work functions of the resulting dual metal gates after annealing are optimal for a dual metal gate CMOS. These criteria are all included in block 11 of Fig. 1. A preferred buffer material is non-stoichiometric aluminum nitride (AlN_x) where "x" is in the range of 0.98 to 1.02. A preferred buffer thickness is less than 1.5nm. Thicknesses larger than 1.5nm of AlN_x may not completely alloy with the gate metal during anneal, and therefore are not preferred. In one embodiment, a gate metal has a thickness of 100nm, a gate dielectric has a thickness of 3.5 nm, and AlN_x, where "x" is near 1.0, has an initial thickness of approximately 1.0nm. Other materials that meet the required criteria are also included in the spirit of the present invention.

It is highly unusual and unexpected that an insulator such as AlN_x can be alloyed with a metal such as Hf to form a metallic alloy with optimal work functions for CMOS applications.

An AlN_x buffer has a very high chemical resistance against chemical etching. Furthermore it can be react with gate metals that have an electronegativity below 1.34, such as Ti (1.32), Hf (1.23) and Ta (1.33), to form alloys that raise the work function of these metals. For example a Hf-AlN_x alloy has a work function of 4.4eV, that is optimal for NMOS and a Ta-AlNx alloy has a work function of 4.9eV, that is optimal for PMOS.

Referring back to Fig. 1, after depositing the buffer layer, a first metal is deposited over the buffer layer, which notably covers first and second wells (block 12). The first metal is then removed from over the second well (block 14). A preferred first metal is Hf. A preferred method of removal of the unmasked gate metal is by wet chemical etching in solutions known to the state of art, including sulfuric acid and hydrogen peroxide, or a mixture of hydrofluoric acid, hydrogen peroxide and de-ionized water. A second metal is then deposited over the first metal and on the exposed buffer (block 16). A preferred second metal is Ta. After two different metals are deposited, etching using a dry etch process such as RIE is done to obtain gate electrode patterns. (Block 18). This CMOS structure including the remaining first and second metal and underlying buffer material, is then subjected to an anneal, whose temperature and time is selected to cause the consumption of the buffer layer by reacting with the gate metal and therefore forming a metal alloy composed of the buffer material and gate metal (block 20). The anneal temperature should be in the range of 400 to 700°C with a preferred temperature of 420°C. The selection of gate metal in combination with the selection of buffer material in the process of the method of the present invention, allows control i.e. a determination of the work function of the metal gate electrode. A particular composition ratio of aluminum and nitrogen can be selected in order to determine the resultant gate metal work function i.e., the work function is dependent on the composition

ratio of aluminum to nitrogen as well as on the anneal temperature and time. Annealing is conducted in a furnace at 400 ~ 500C for 30 mins, or in RTA tool at 500 to 700C for 1 min.

The process/method of the present invention will now be described more clearly in reference to Figs. 2A-2G. Fig. 2A shows a prior art substrate 22 (preferably $S_{i,}$), a gate dielectric 26, and symbolically indicates an NMOS p-well 23 and a PMOS n-well 25. Those skilled in the art will understand how to construct Fig. 2A and variations, all of which are included in the present invention in combination with the gate structure and method of construction as described in reference to Fig. 1 and described in detail below.

According to the present invention as shown in Fig. 2B, a buffer layer 28 is deposited on the gate dielectric 26. The buffer layer 28 prevents the gate dielectric/gate oxide 26 from being exposed to the metal etching process, and also determines the work functions at the metal/dielectric interface. During the annealing, the buffer layer is completely consumed through reaction with the gate metal, and a new alloy is formed. This process has the additional advantage that there is minimal change in the equivalent oxide thickness of the gate dielectric region. The buffer layer material is selected to meet the requirements discussed in reference to block 11 of Fig. 1. The preferred embodiment of the present invention includes a buffer material of AlN_x, with "x" approximately 1.0 in the range of 0.98 to 1.02. One of the buffer material requirements discussed above is that it must convert with the gate metal into a metal alloy during the annealing process. AlN_x can be converted in the annealing process into metal alloys when reacting with metals that have electronegativity below 1.34, such as Ti(1.32), Hf(1.23) and Ta(1.33) to form alloys with altered work functions. Alloys of these metals with AlN_x have work functions substantially higher than that of the bare metals.

Fig. 2B also illustrates the next step in the process/method wherein a first gate metal 30 is deposited on the buffer layer 28. Metal 30 in the example given is an NMOS

metal. This first gate metal could alternatively be a PMOS metal such as Ta (not Hf). Assuming the first metal 30 is an NMOS metal, the NMOS metal must then be removed from over the PMOSFET region. Alternatively, if the first metal were a PMOS metal, it would have to be removed from the NMOSFET region. In the example given, a photoresist 31 is placed over the first metal 30 (NMOSmetal) as shown over the NMOSFET region, and the metal is etched away from over the PMOSFET region, resulting in the structure as illustrated in Fig. 2C. With the NMOS metal removed from the PMOSFET REGION, a PMOS metal 32 is deposited as shown in Fig. 2D, covering the NMOS metal 30 and the PMOSFET area. This is shown planarized in Fig. 2D. The next step is to remove both the PMOSFET and NMOSFET metals, and the buffer layer, except in the NMOS and PMOS gate areas, symbolically indicated as 34 and 36 respectively in Fig. 2E. Those skilled in the art will know of various methods of accomplishing this removal. The method illustrated simply places resist 38 over the two gate areas as shown in Fig. 2E, and etches the remaining exposed metals. The etching of the second metal is different from the first metal etching. The first metal is etched selectively while the second metal etching is only to define the gate electrode pattern. The buffer layer 28 remains after etching as shown in Fig. 2F. However, please note that the buffer layer is consumed by the annealing process, forming alloys 38, 40 with the deposited metal as shown in Fig. 26.

The present invention also includes other methods of achieving the structures of Fig. 2F that use the novel buffer layer 28 for the purpose as disclosed above. The next step is to anneal the structure of Fig 2F for alloying the buffer layer with the metal layers 30 and 32 for the NMOSFET and PMOSFET gates respectively to consume the buffer layer and form metal alloys, indicated as 38 and 40 in Fig 2G. The buffer layer 28 has effectively been consumed in the annealing/alloying process.

Fig. 3 is presented to show the etching rates of various films. HPM is a mixture of HF, H_2O_2 and H_2O, and SPM is a mixture of H_2SO_4 and H_2O_2. As shown, HPM has a very low etch rate on AlN_x compared with Hf or S_iO_2. SPM has zero etch rate on AlN_x , compared with zero etch rate in S_iO_2 and a very high etch rate on Hf.

Fig. 4 includes a curve "-o-" representing the gate capacitance vs. gate voltage of a gate structure having Pt gate metal deposited directly on a S_iO_2 gate dielectric, without a buffer layer. The curve "-●-" represents the gate capacitance vs. gate voltage of a gate structure with a Pt gate metal on an AlN_x buffer on a S_iO_2 gate dielectric after anneal at 420°C. The structure of curve "-●-" was not exposed to an etchant for removing for example Hf metal, and therefore sets a reference for comparison. The curves indicate a capacitance at -2V of approximately 820 nF/cm^2 for Pt/S_iO_2 and 750nF/cm^2 for the $Pt/AlN_x/S_i0_2$ (annealed), which corresponds to a difference in equivalent oxide thickness of 0.3 nm, assuming equal dielectric constants for AlN_x and S_iO_2.

The "-▲-" curve represents the gate capacitance vs. gate voltage for a $Pt/AlN_x/S_i0_2$ structure after anneal, but in the construction process was subjected to a wet chemical etch on the AlN_x layer prior to deposition of the Pt layer. This was done to test the effectiveness of the AlN_x layer in resisting the wet etch that is used to remove the portion of the first metal 30. The curves show no significant difference between the "-●-" curve and the "-▲-" curve, indicating that the AlN_x was effective as an etch mask in resisting the wet chemical Hf strip process.

As a further test/evaluation, a $Ta/AlN_x/S_i0_2/S_i$ gate layer was formed and a High Resolution Transmission Electron Microscopy (HRTEM) image was taken showing the thickness of the combined AlN_x/S_iO_2 layers before annealing, and after annealing at 420°C. The thickness was 4.24nm prior to annealing, and 3.50nm after annealing. The difference of 0.74nm in thickness before and after anneal, as shown in TEM pictures, showed the

consumed AlN. This confirmed the consumption of AlN. It should be noted that the value of 0.74nm is below 1.5nm.

Fig. 5 shows curves of gate capacitance versus gate voltage for two structures, one using hafnium (Hf) and the other tantalum (Ta) gate metal. The curves are after a 420°C anneal, and indicate a maximum difference in the curve of 0.5V.

Fig. 6 is a plot of the gate forward bias voltage (V_{fb}) versus the gate oxide thickness (T_{0x}) for two different gate structures after anneal, both using the AlN_x buffer layer in the process according to the present invention. The evaluated work functions are 4.9eV with Ta as the gate metal, and 4.4eV with Hf (hafnium) as the gate metal, again showing $\blacktriangle Ø$=0.5V.

Fig. 7 is a plot of the change in the equivalent oxide thickness resulting from the annealing process for various annealing temperatures and for two different structures, one with a AlN_x thickness of about 0.8nm and the other with a AlN_x thickness of about 1.5nm. The graphs show that the equivalent oxide thickness is reduced slightly as a result of the annealing process and reduced more for the thicker structure. The maximum change was about 0.5nm corresponding to a 700°C anneal with use of the 1.5nm AlN_x layer in a $Ta/AlN_x/S_iO_2$ stack.

Although the present invention has been described above in terms of a specific embodiment, it is anticipated that alterations and modifications thereof will no doubt become apparent to those skilled in the art. It is therefore intended that the following claims be interpreted as covering all such alterations and modifications as fall within the true spirit and scope of the invention.

What is claimed is: